Recipes for Safety in Open-domain Chatbots

Year: 2020 | Citations: 243 | Authors: Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, J. Weston

---

Abstract

Models trained on large unlabeled corpora of human interactions will learn patterns and mimic behaviors therein, which include offensive or otherwise toxic behavior and unwanted biases. We investigate a variety of methods to mitigate these issues in the context of open-domain generative dialogue models. We introduce a new human-and-model-in-the-loop framework for both training safer models and for evaluating them, as well as a novel method to distill safety considerations inside generative models without the use of an external classifier at deployment time. We conduct experiments comparing these methods and find our new techniques are (i) safer than existing models as measured by automatic and human evaluations while (ii) maintaining usability metrics such as engagingness relative to the state of the art. We then discuss the limitations of this work by analyzing failure cases of our models.